Basic Concepts in the Study of Diseases with Complex Genetics

Margit Burmeister

Most diseases run in families—this is also true of virtually all psychiatric disorders. Twin and adoption studies have shown that most psychiatric disorders have a genetic component, yet very few genetic factors are known, as is true for most disorders with a complex genetic origin. Here I review, for nongeneticists, some of the basic terminology and concepts used when studying complex genetic diseases, with examples from psychiatric genetics. This review is intended to help in the understanding and critical evaluation of reports on genetics of psychiatric illnesses in the literature. Biol Psychiatry 1999;45: 522–532 © *1999 Society of Biological Psychiatry*

Key Words: Complex genetics, psychiatric genetics, association, linkage

How Do We Know That Predisposition to a Disease Is Genetic?

Most diseases run in families—but that is not enough to conclude that genetic factors are involved, since infectious diseases or nongenetic traits such as malnutrition or attendance of medical school also run in families. Twin and adoption studies discriminate between familiality due to genetic or due to environmental influences (geneticists mean with the latter any nongenetic factor, including chance and measurement error). Adoption studies have shown that the risk of an adoptee of having a psychiatric disorder depends more significantly on the mental health of the biological parent than on that of the adoptive parent (Kendler and Diehl 1993; Mitchell et al 1993).

In twin studies, concordance rates among monozygotic and dizygotic twin pairs are compared. To measure concordance rates we simply ask: if one twin is affected, what is the chance that the cotwin is also affected? If both monozygotic and dizygotic twin pairs have similar high concordance rates, a similar (shared) environment is a major factor for susceptibility. Similar low concordance rates are due to different (nonshared) environmental influences. A genetic influence is indicated if monozygotic are significantly higher than dizygotic concordance rates. If only one gene contributes in each family, we expect monozygotic twins to have a twofold higher concordance rate than dizygotic twins, since they share twice as many genes as dizygotic twins. If more than one genetic factor contributes to susceptibility in the same person (which can be two or more different genes, or, as in a recessive disorder, two alleles of the same gene), the difference between the concordance rate for monozygotic versus dizygotic twins is expected to be larger than twofold. Such a contribution of multiple genes to a given illness is often called epistatic gene interaction (Lander and Schork 1994; Plomin et al 1994). Epistatic interaction should be contrasted with heterogeneity (see below), in which several genes can cause the same disease, with only one contributing in each individual. Twin studies cannot detect whether there is heterogeneity, and would still implicate one gene, since only one gene acts in each family.

Twin data implicate genetic factors as major predisposing factors for most psychiatric illnesses (Plomin et al 1994). The concordance rates indicate a stronger genetic influence than for many other complex disorders, although often the difference between monozygotic and dizygotic twins is less than a factor of two, indicating both genetic and shared environmental factors (Plomin et al 1994). For schizophrenia and autism, there is also evidence from twin and family data for epistatic interaction, i.e., several genes interacting in the same individual (Plomin et al 1994; Risch 1990). [Also, see Report of the National Institute of Mental Health's Genetics Workgroup (Appendix E) summary of genetic findings in all psychiatric disorders elsewhere in this issue.]

Alleles—The Variations in the Genome

In general, an allele is one of several forms of any defined DNA sequence in the genome, whether it is a gene or an anonymous DNA sequence. Such a defined place in the genome is called a locus [plural loci, please see also the glossary (Appendix 1) for a summary of terms explained in this section]. If a disease is inherited in a simple manner, the inheritance of disease alleles can be followed in pedigrees.

Although any base pair change is caused molecularly by mutation, the word mutation is now typically used in a

From the Mental Health Research Institute, University of Michigan, Ann Arbor, Michigan.

Address reprint requests to Margit Burmeister, PhD, Mental Health Research Institute, University of Michigan, 205 Zina Pitcher Place, Ann Arbor, MI 48109-0720.

Received July 28, 1998; revised August 14, 1998; revised August 24, 1998; accepted September 2, 1998.

0006-3223/99/$19.00
PII S0006-3223(98)00316-3

more restricted sense, implying disease-causing mutation. An allele that is frequent (>1%) in the general population is called a polymorphism. With polymorphisms found every few hundred base pairs in a recently sequenced example (Nickerson et al 1998), and predicted to exist in abundance in the human genome (Chakravarti 1998; Wang et al 1998), it is becoming increasingly clear that the study of genetic variation will be part of the Human Genome Project, which will provide us in the near future with hundred of thousands of polymorphisms spread over all of the genome (Chakravarti 1998; Collins et al 1997; Wang et al 1998).

Large-scale genetic mapping became possible when it was realized that variants are frequent in the human genome, and molecular tools such as the polymerase chain reaction (PCR) became available to analyze them in large numbers of samples. Polymorphisms used in genetic studies often are called (genetic) markers. The first large-scale human genetic mapping studies used variants that affect restriction sites, termed restriction fragment length polymorphisms (RFLPs). A restriction enzyme cleaves at a specific sequence. For example, EcoRI cleaves at the sequence GAATTC. RFLPs arise when such a site is mutated, e.g., to GAATTT, which is no longer cleaved by EcoRI. Most RFLPs only have two alleles, one with the restriction site, which is cleavable by the restriction enzyme, and one allele without the site. When following such alleles in families, an individual will often have the same allele on both chromosomes, and thus inheritance of these alleles cannot be followed in a pedigree, and this branch of the pedigree is termed noninformative.

The real boost for genetic mapping came when DNA sequences were discovered with many more alleles, usually between five and 10 different alleles at each locus, called microsatellite markers, simple sequence length polymorphisms (SSLPs), or short tandem repeats (STRs). Here I will use the term SSLPs. As the name suggests, these are simple sequences that are repeated a variable number of times; for example, a single locus may have five different alleles, each consisting of the simple sequence GT repeated 17, 18, 19, 21, or 24 times, resulting in length differences of two or more base pairs. These repeats are flanked by unique DNA sequences for which PCR primers can be developed, and the length difference is detected on gels. This process can be performed in a fairly automated fashion on fluorescent sequencers, and requires only small amounts of DNA, allowing the large number of studies currently being performed. The most common SSLPs repeat the sequence GT (CA on the opposite strand), although for technical reasons many genome scans now prefer to use 4-bp repeats such as (GATA)n (Murray et al 1994). There are over 6000 known SSLPs, with known sequence, primers, PCR conditions, and precise chromosomal location (Dib et al 1996; Murray et al 1994).

More recently, several thousand simple single nucleotide (base) changes (single nucleotide polymorphisms or SNPs) have been identified (Wang et al 1998). Like RFLPs, there are usually only two alleles at each locus; however, more than the gel-based technology of SSLPs, these may become more readily automatable in the near future, using new procedures involving hybridizations to "chips" (see Watson and Akil current issue). The ability to analyze thousands of SNPs more efficiently than SSLPs compensates for the higher likelihood that any one SNP marker may be uninformative.

Diseases Can Have a Complex Genetics for Many Reasons

A Mendelian disease runs in families in a strict dominant, recessive, or X-linked fashion. Hundreds of such disease loci have been mapped, and over 600 genes involved in genetic diseases have already been identified (Gelehrter et al 1998); however, so far there are only very few examples of psychiatric illnesses inherited in a strictly Mendelian fashion (see Brunner et al 1993 for such an example). Since there is some variation in onset and clinical course even in strictly Mendelian disorders such as cystic fibrosis, what makes a disease complex, i.e., non-Mendelian? Lander and Schork (1994) list the following basic problems: a) incomplete penetrance, i.e., someone who carries the disease allele may not become ill, or the onset may be extremely late; b) phenocopy, i.e., someone, even with relatives with the genetic form, may be ill for a nongenetic reason; c) heterogeneity, i.e., mutations in many different genes can have the same clinical end result; d) polygenic inheritance, for example the additive effects of several different alleles on quantitative traits such as blood pressure, or epistatic interaction, as suspected in schizophrenia and autism, where several predisposing alleles have to come together; e) high frequency of the predisposing alleles and of the disorder; and f) other genetic mechanism of inheritance, such as mitochondrial inheritance, or a genome that is actively changing, as in disorders with trinucleotide expansions.

Mapping and Cloning of Mendelian Disease Genes: Parametric Linkage Analysis

Before going into the problems of complex disorders, let us review how Mendelian disease genes were mapped so successfully in the past 10 years (Collins 1995). To identify such genes, families are ascertained, preferentially large pedigrees, DNA is isolated from blood, and linkage analysis is performed in which SSLPs from all over the

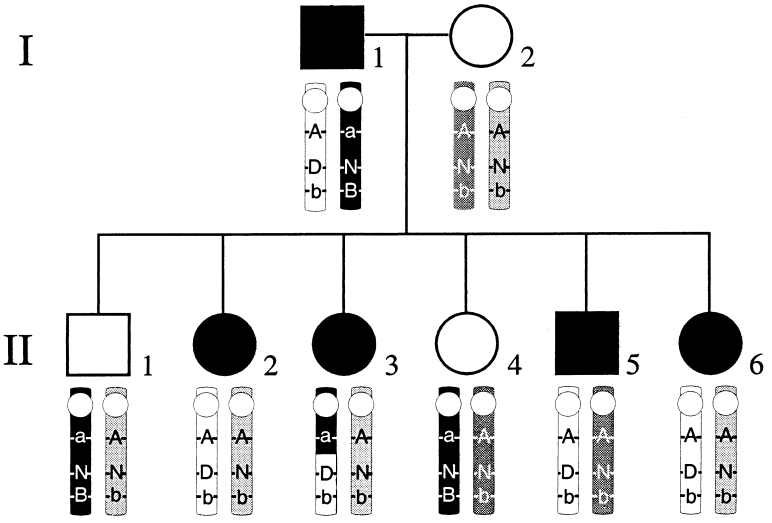

Figure 1. Linkage and recombination. Individuals shown in solid symbols are affected with a dominant disease due to the presence of the disease allele D, N being the normal allele. The father carries allele D on a chromosome with marker alleles A and b. Most affected offspring inherited also the A and b alleles from the father because these alleles are linked to the disease-causing D allele on the same chromosome. In this example, marker allele b is depicted as being closely linked with D, whereas marker allele A is somewhat more distantly linked. Thus, individual II-3 demonstrates a recombination event. During meiosis in the father, allele A was separated from the disease allele D, resulting in inheritance of the a allele from the father. Allele b did not recombine with the disease allele D.

genome are tested—i.e., PCR reactions are performed on DNA from each member of the pedigree with 200–500 SSLP markers. The logic of linkage analysis is straightforward: If a disease-causing dominant mutation is, for example, in a gene in the middle of chromosome 4, all affected members in a pedigree should receive the exact same region of chromosome 4 around that mutant allele (see Figure 1). Obviously, that part of chromosome 4 carries also other DNA, which is physically "linked" on the chromosome to the mutation, among it some DNA that contains SSLP marker loci. That specific chromosome with the mutation carries only one particular allele of each SSLP marker, since all affected members in a family came from a common ancestor. In that case, the common alleles are "identical by descent," meaning identical because they descended from the exact same chromosome. If an SSLP marker is closely linked, i.e., near the mutation, affected individuals in a family are expected to have inherited the same allele from an affected parent (allele b in Figure 1), whereas unaffected members will receive different alleles from their parents.

If the marker is on the same chromosome but a certain distance away (see alleles A and a in Figure 1), however, the marker allele will sometimes be separated from the disease gene during meiotic recombination (individual II-3 in Figure 1). The further away the marker locus is from the disease-causing mutation, the more often this happens. Thus, the rate of recombination is a measure of distance on the chromosome between the SSLP marker and the disease-causing mutation. If there is a break between two loci on average in 1% of all meioses, they are said to be 1 cM apart—which roughly correlates with 1 million base pairs distance between the two markers. Because these linked SSLP alleles generally do not cause the disease and just "mark" the chromosome, each family can have a different allele of the SSLP linked to the disease.

Once we have found an allele of a marker locus in a pedigree that segregates with the disease, how do we find out whether we have enough information? If we would look at one family with four children, two of whom are affected, chances are we may find several alleles on different chromosomes segregating with the disease by chance. To find out how certain we can be that a marker is linked, a statistical analysis is performed; we compare the likelihood of getting the specific constellation of marker alleles and affectation status in the specific family under two hypotheses—the "test" hypothesis that the marker is linked to the disease at a specific distance, reflecting the recombination rate, called θ (theta), under a specific model of genetic transmission (e.g., dominant), compared to the likelihood of these results under the neutral ("null") hypothesis that there is no linkage between the marker and the disease. Dividing these two likelihoods gives a quotient, called the likelihood ratio. The logarithm to base 10 of that ratio is the LOD (logarithm of odds) score. The LOD score can be calculated by modeling in our test hypothesis any distance θ of the marker to the disease locus. Once linkage is significant, the point with the highest LOD score gives the most likely distance between the marker and the disease locus. Traditionally, a LOD score of 3, corresponding to an odds ratio of 1000:1, is accepted as evidence for linkage; however, a confusing point is that a LOD score of 3 is not equivalent to $p = .001$, i.e., it does not mean that the chance that we are

wrong with our test hypothesis is only 0.1%. The reason is that when we performed a genome scan to find linkage, we have performed multiple tests, with markers all over the genome, with each one of these markers a priori being quite unlikely to be linked. In fact, with a LOD score of 3.0, our chance to be wrong is about 9%—assuming a simple Mendelian disease (Lander and Schork 1994). To model the traditional threshold of claiming significance, i.e., $p = .05$, a LOD score of 3.3 needs to be achieved (Lander and Schork 1994).

One advantage of the LOD score is that studies can be compared and analyzed together very simply; if one investigator found tentative linkage for a disease to a specific marker with a LOD score of 1.5, and another investigator, studying a different family with that same marker, also found tentative linkage, with a LOD score of 2.0, the LOD scores can simply be added (assuming both used the same diagnostic criteria and genetic model). Although neither study has strong enough data to be sure of linkage, both studies together are significant with a LOD score of 3.5.

To increase the power and certainty with which linkage is detected, multipoint linkage analysis is often performed. This simply means that several marker loci, which are known from other studies to be close to each other, are simultaneously checked for linkage with the disease in question. If there is true linkage, all markers adjacent to the disease locus are expected to be linked. The combination of alleles that segregates together in a family with the disease is also called the linked haplotype, a hapolotype being a combination of alleles for a number of loci on the same chromosome that segregate together, in this case in a family. The concept of haplotypes is also useful in population studies, when alleles segregating together in a population are considered, as discussed below.

Once we know the location of a disease gene, more families linked to this same locus are examined, to narrow down a small region on the chromosome where the gene is located. If that is feasible, the steps that follow, namely cloning of the region and identifying every gene within the linked region, albeit hard work, have become easier and more straightforward (Collins 1995). With the advances of the Human Genome project, it will be even easier, since one may soon simply look at the already existing sequences in the linked region for candidate genes.

A major problem when studying complex diseases such as mental illness is that the linkage analysis described above is parametric, i.e., model based. The two hypotheses that are compared (test hypothesis linked at a specific distance and null hypothesis of no linkage) are very specific. Thus, we have to specify dominant versus recessive or another mode of transmission model—and for most complex disorders we do not know that for certain, and it may be different in different families; we also have to specify for every person in the pedigree whether or not he or she is affected, which is difficult when you have, for example, individuals with unipolar depression in a bipolar or schizophrenia pedigree. Sometimes, a specific model of penetrance is modeled on the pedigree and marker data, taking into account the typical age of onset of the disease and the age of each individual. The solution to the question of what model to use is usually to model many different hypotheses in a linkage study, and to report the one giving the highest LOD score; however, the drawback then is that we have to account statistically for the many hypotheses that were tested. Thus, the LOD score to even be reasonably sure to have found linkage has to be much higher than the traditional threshold (Lander and Kruglyak 1995; Lander and Schork 1994). So far, very few linkage studies on major psychiatric disorders have yielded consistent linkage results, but a few loci, on 6p for schizophrenia and 18 for bipolar disorder, have been found in at least two large independent studies (Gershon et al 1998).

Heterogeneity

One complication in linkage mapping is if a disease runs in a Mendelian fashion in families, but in each family a different gene, usually on a different chromosome, causes the disease. This is called heterogeneity. One prime example is deafness, for which so far over 30 different loci have been identified (Petit 1996). Deafness can be inherited in a recessive, dominant, or X-linked fashion, and each runs in families in a fairly Mendelian fashion, so that it seems often (but not always) to be a Mendelian disease. Although combining data from several families with different predisposing loci is possible if certain conditions are met (Risch 1989), linkage for very heterogeneous disorders such as deafness is usually established in single large families studied one family at a time, or in remote areas of the world where only one deafness gene is present and the population effectively presents a large extended pedigree (Friedman et al 1995). This approach is being used for bipolar disorder in studies of the Costa Rican and the Amish population (Freimer et al 1996; Ginns et al 1996). But again, one has to be careful; with a disorder as common as bipolar disorder, in a large pedigree, chances are high that individuals carrying a different predisposing gene marry into the pedigree—even if they themselves are not affected. Then the investigator may in fact be looking at linkage of two or more different loci at the same time, making linkage analysis much more difficult (Pauls et al 1995). Alternatively one can try to find a clinical form that stands out as a special genetic form, as was the case for early-onset Alzheimer's disease (Roses 1998).

Nonparametric Linkage Analysis

To avoid some of the problems of selecting a very specific model, while still using some of the power of linkage analysis, nonparametric, i.e., mode-of-inheritance-independent, methods of linkage analysis were developed, called affected sib pair (ASP) methods or the more general affected pedigree member method (APM) (Weeks and Lange 1988). In these methods, only sibpairs or other pairs of affected relatives are studied, which means the power of seeing alleles segregating in large pedigrees is lost. The idea of these methods is that independent of whether the disease is dominant, recessive, or more complex, if there is a disease-causing mutation in a specific chromosomal region in a high proportion of families, we expect two affected individuals from the same family to share an allele of a marker locus more often than expected by chance alone (50% for siblings). Just like in parametric linkage, which marker allele is shared is expected to differ from family to family, so what is scored is only how many alleles are shared, i.e., 0, 1, or 2. One can design the study to include only pairs with a specific diagnosis (e.g., only bipolar I disorder, excluding bipolar II, major depression, and other disorders), avoiding unclear phenotypes. In addition, no genetic model needs to specified, and linkage can be detected in the presence of heterogeneity (Weeks and Harby 1995). The statistical analysis is also simpler. Terms important in APM studies are identity-by-descent (IBD) and identity-by-state (IBS). These can be illustrated on the simplest case, two affected siblings. If both carry an A and a B allele, the two siblings are called IBS for both alleles at that locus. They share two alleles by state, which means both alleles are the same length whatever the measurement of the alleles was, but we cannot say where they came from; however, if both parents are also AB heterozygotes, we cannot actually conclude that they received the same chromosomal regions from their parents —they may in fact share none. Thus, the two affected siblings may be IBS (both AB) but not IBD (the two sets of A and B alleles came from different parental chromosomes). If, however, the two parents are AC and BC heterozygotes, we are sure the two affected siblings with AB alleles received the same alleles from the same chromosomes; therefore they are also IBD. Thus, knowing the parental genotype makes APM more powerful, since we are interested in linkage to a chromosomal region, and can conclude for certain whether there is IBD or not with parental genotypes.

It is not always possible to get the parental genotype, however, as for example in studies on Alzheimer's disorder (Roses 1998). Since we are not really interested in IBS, but in IBD, we would like to put the most weight on those cases in which IBS is most likely a reflection of IBD. That is done by taking the allele frequencies of each marker into account: if, as in the example of the two siblings both being AB for a marker, A has a frequency in the general population of only 2%, but B has a frequency of 60%, we will put more weight in our statistical analysis on the fact that allele A is shared; it is less likely to be present in both parents, and thus it is more likely that allele A is indeed shared IBD, i.e., that it is derived from the same parental chromosome (Weeks and Lange 1988). This consideration, however, requires knowledge of the allele frequencies in the population studied, and can give misleading results if the population or the allele frequencies are misspecified.

Because of its robustness, the APM approach is often used as the first approach to identify linkage, even when extended pedigrees for linkage studies were initially collected, and a genome scan of markers can be performed on pairs within these larger pedigrees (Brown et al 1994). APM was successful in identifying chromosome 19 for late-onset Alzheimer's disease (Roses 1998). Currently, for many complex diseases APM analysis is performed, as was the case for example in the recently published large linkage studies of alcoholism (Reich et al 1998).

Case–Control Association Studies

Linkage-based studies are based on following marker alleles that are close to a mutation on a specific chromosomal segment. These approaches do not make any biological assumptions about the disease, in fact they are performed in exactly the same fashion for diabetes, hypertension, or alcoholism. Once a location is found, the nature of the linked marker is irrelevant, all it does is "mark" the region of the chromosome. Popular press releases often confuse identification of linkage with finding a gene—finding linkage only means we have a reasonable likelihood of knowing where on a chromosome to look for the gene!

But can a susceptibility gene be identified in a more directed manner, incorporating our knowledge of the disease? Association studies do just that, and have been popular for that reason for many years. This approach has often been called candidate gene approach, although geneticists tend to refer to a candidate gene as a gene that is located in a chromosomal region previously found relevant in linkage studies (Gelehrter et al 1998). An association study is basically a genetic case–control study (Figure 2A). The logic relies on our hypothesis about the illness in question; we ask, for example, if a functionally different allele of the serotonin transporter is more frequently observed in depressed or in alcoholic patients than in control subjects. The methods and statistics involved superficially are very simple; we ask if the patient sample

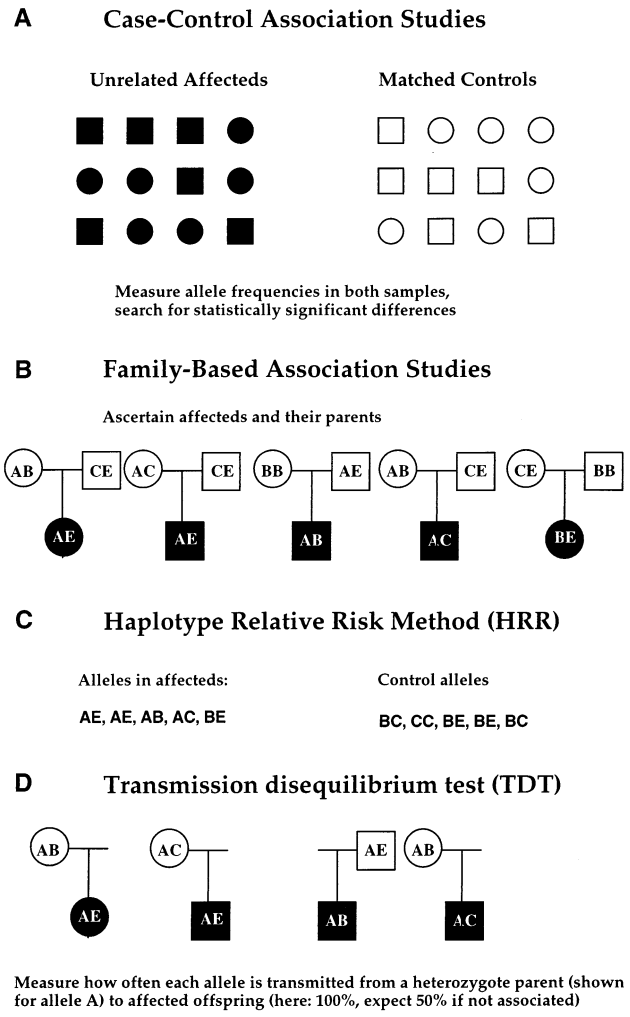

Figure 2. Association studies. **(A)** Case–control association studies compare allele frequencies between a patient cohort and an ethnically matched control cohort. The problem here is that ethnic matching is not completely feasible. **(B)** Family-based association studies overcome the problem of ethnic matching by ascertaining patients as well as their parents. Nontransmitted alleles from parents can be used as controls. **(C)** In the haplotype-relative risk method, the allele from each parent that is not present in the patient is used as a control. **(D)** In the transmission disequilibrium test, each allele, in this example allele A, is tested separately. Only parents heterozygous for allele A are relevant. A heterozygous parent is expected to give allele A to 50% of his/her offspring; however, if allele A is a predisposing factor, it is transmitted more often than expected to affected offspring (4/4 times =100%, in this small example).

has more of one allele than the control sample. Alleles known or presumed to have a functional significance because they affect the protein coding region or the level of expression are currently the most interesting to evaluate. Since many psychiatric disorders are frequent, and we have to assume that there is epistatic gene interaction, the predisposing alleles are also expected to be fairly common polymorphisms rather than rare mutations, as is the case for Mendelian diseases. Examples of interest for mental illness are alleles of the dopamine D4 receptor with different C-terminal regions due to a protein-coding repeat domain of different length (Van Tol et al 1992); catechol-O-methyltransferase variants with low and high activity (Lachman et al 1996); and serotonin transporter alleles with a high or low activity promoter (Lesch et al 1996; see also Malhotra and Goldman, current issue).

Many association studies have resulted in controversial, irreproducible, or erroneous results, however, as exemplified by a widely publicized study claiming that an RFLP within an intron of the dopamine D2 receptor gene (DRD2) is associated with alcoholism (Blum et al 1990). This association could not be reproduced by most laboratories, and most now agree that this allele of DRD2 is no longer considered a major risk factor for alcoholism (Gelernter et al 1993). This does not necessarily exclude the DRD2 gene, since other alleles of the same gene are not necessarily excluded.

As anyone performing clinical trials knows, case–control studies rely on a design in which the control subjects are comparable to the cases, which is why complicated formulas are used to match patients on gender, age, risk, weight, and other parameters. The problem with case–control studies in genetics is that it is hard to define and ascertain a precisely genetically matched control sample. In the case of the RFLP in DRD2, later studies showed that the frequency of the allele in question can vary widely among populations, between 9 and 75% (Barr and Kidd 1993). Thus, if the ethnicity of the cases was slightly different from the ethnicity of the control subjects, perhaps because the incidence of alcoholism varies among ethnic groups, allele frequencies are expected to differ, without any relevance for alcoholism. Ethnicity here is not limited to broad categories—of course, an association study with Caucasian samples uses Caucasian control subjects—but it may be quite subtle, such as differences between Norwegians and Finns. An illustrative example is a study of diabetes in Native Americans (Knowler et al 1988); since the study was limited to a specific Native American tribe in a reservation, one would think that ethnicity was well controlled. A strong association was identified in which one particular allele of the immunoglobulin complex seemed to protect from diabetes; however, thorough analysis in which participants were asked about the ancestry of all great-grandparents showed that the associated allele was in fact a marker for white admixture—since diabetes is less common in Caucasians than in the tribe studied, having some Caucasian ancestry protected from diabetes, and the

Table 1. Comparison of Different Methods Currently Used in Complex Genetics

	Linkage studies		Association studies	
	Parametric	Nonparametric (ASP/APM)	Case–Control	Family-based (HRR and TDT)
Power	High for Mendelian, low for complex	Moderate for Mendelian and large effect complex	Low for Mendelian, higher for alleles with small effects	Highest for alleles with small effects, low for Mendelian
Patient resources needed	Large pedigrees with many affected subjects with same defect	Two affected subjects per family, best when parents also available	single patients and ethnically matched control subjects	Patients and their parents, preferentially also other sibs
Genetic tools needed	Genetic markers (SSLPs or SNPs) every 5–20 cM	Genetic markers (SSLPs or SNPs) every 5–20 cM	Candidate genes with polymorphic alleles	Candidate genes with polymorphic alleles
Advantages	Highest power if Mendelian, genome scan very efficient	Highest power if large gene effect; no genetic model needed; allows genome scan	Ease of sample collection and statistical analysis	Highest power to find alleles with small effect on risk
Problems	Need to ascertain pedigrees; need specific genetic model; limited if heterogeneous	Need parents or siblings for best power (IBD); not easy to narrow down to identify gene	Cases and control subjects have to be ethnically well matched to avoid false positive association	Need to ascertain large numbers of patients and parents; need candidate gene
Summary	Best applied to Mendelian diseases with no or limited heterogeneity	Best for first pass genome scan of somewhat complex diseases and to identify candidate genes	Easiest to check if a candidate gene might be involved, but prone to false positives	Best for highly complex disorders, but needs alleles in candidate genes.

Please refer to text and further literature, since many statements made here are correct only under certain conditions. For example, it can be anticipated that genome-wide association studies, without specifying candidate genes, may be possible in the not too far future.

allele is more common in Caucasians. Since in the United States the ethnic origin is often very mixed, and we do not know a priori the ethnic distribution with precision, such an artifact, called population stratification, cannot easily be avoided. It is especially a problem when allele frequencies, as in the case of DRD2, vary widely across populations.

Because ethnic stratification is not due to small sample size but due to unequal distribution of alleles, increasing the sample size will not eliminate this problem, it will even make the apparent association stronger. A way to partially overcome this problem is ascertainment of cases and control subjects from many different populations. If the same allele is associated with the disease in many different populations, we can be more certain (but not absolutely sure) that the allele is truly associated with the disorder.

Family-Based Association Studies

A better way to overcome the population-stratification problem in case–control studies, however, are newer, family-based approaches, called haplotype relative risk (HRR) and transmission disequilibrium test (TDT) (see Figure 2B–D). The idea in these studies is to use the nontransmitted allele from the parents of an affected proband as internal controls (Figure 2B). In this manner, the same individuals, parents of affected subjects, provide both the test and the control sample, and thus the issue of ethnic matching is avoided. Recent research also suggests that collecting unaffected siblings will increase the power of certain studies, especially when dealing with quantitative traits (Risch and Zhang 1995).

In HRR (Falk and Rubinstein 1987; Terwilliger and Ott 1992), the alleles from each parent that are not transmitted to the patient are used as control samples (Figure 2C); if one parent carries alleles A and C, the other A and B, and the patient carries alleles A and B, one A and one C allele were not transmitted to the patient, and are the control sample.

The basis for TDT is an apparently skewed transmission of a predisposing allele; if a parent is heterozygous for allele A, we expect on average 50% of the offspring to inherit allele A, and 50% the other allele. If, however, allele A is a predisposing allele for a disorder, we expect patients to receive allele A more often than expected by chance (Figure 2D). TDT analysis (Spielman and Ewens 1996) will only yield a positive result when the allele is both associated with the disorder and linked—which is what we expect from an allele predisposing to the disease but not from an allele that marks ethnicity.

Which Method Might Work for Which Disease?

How can we decide between the methods mentioned above for our particular interest (see Table 1 for a

simplified summary)? When the disease is inherited in a clearly Mendelian fashion, there is no question that parametric linkage analysis is by far the most powerful method to locate a gene. Several pedigrees have been identified in which bipolar disorder segregates in an apparently Mendelian dominant fashion, and other pedigrees in which it seems X-linked; however, since a disease such as bipolar disorder, which clearly runs in families, is fairly common, some families with an apparently Mendelian inheritance pattern are expected to be found by chance (Hebebrand 1992). Such families will not give consistent results in linkage studies. In contrast, early-onset Alzheimer's disease (Roses 1998) is inherited in a dominant manner, and thus was recognized as a special, clinically distinct form of an otherwise complex disease, and linkage was applied successfully. Linkage may thus be useful when a specific, clinically unique form of the disease is studied. Another approach may be to search for Mendelian traits associated with psychiatric disorders, e.g., electrophysiological abnormalities, and then perform parametric linkage analysis (see Freedman et al current issue), or to identify clinically distinct families with clearly Mendelian inheritance (Brunner et al 1993).

To determine which approach is most useful, we need to get an idea of how complex the disease is. One way to ask the question is to say: how much of a difference does it make for my risk for a disease if I have a relative with the illness? In other words, if I have a depressed mother, how much greater is my chance of becoming depressed than that of the population at large? This measure is the recurrence risk, lambda (λ), which is defined as the risk of a relative of an affected subject to have the disorder, divided by the risk to the general population (Lander and Schork 1994; Risch 1987, 1990). This value can be calculated from epidemiological data; so, if the risk to a sibling of an autistic proband is about 8%, and the incidence of autism in the general population is about 4/10,000, λ_S, the risk ratio for a sibling, is about 200. For depression, the risk ratio is much lower, because depression is quite prevalent in the general population. In general, the larger the risk ratio, the easier it is to find a disease gene; however, autism may be a proof to the contrary, since there may be complex interaction of several genes with each other, and genes with environmental factors.

Another way is to ask: how is my risk increased if I carry one specific predisposing allele of a specific gene? This is the genotypic relative risk (GRR) (Risch and Merikangas 1996). GRR describes the increased chance of having the disease for an individual with one allele over the risk if he carried the other allele (everything else being imagined to be equal). Even with large λ, the GRR may be small if there are many genes that each contributes a little to susceptibility, making a study very complex, as may be the case for autism. If an allele only increases risk by a factor of 2, the effect of that gene on the risk for the disease is not very strong. GRR is a useful concept for modeling, but we do not usually know the GRR when a study is initiated—it can only be determined once we have identified such a predisposing allele. When various values for GRR were modeled, it became clear that for large GRR, that means alleles with a large effect, linkage studies, including APM, are more powerful and require fewer samples; however, as soon as GRR fell below about 4, the number of samples required for linkage studies became astronomical, yet were still within reasonable range when TDT was modeled as the method of analysis (Risch and Merikangas 1996). Thus, the more complex the disorder, and the smaller the effect size of each allele, the more advantageous it is to use TDT; however, compared to case–control association studies, samples for TDT are harder to obtain; it is much easier to collect patients and control subjects than to collect patients and their parents.

Genome-Wide Association Studies

In any form of linkage study, with a few hundred markers we can search blindly, without biological hypotheses, over the whole genome. In contrast, in TDT, specific alleles of a candidate gene are tested for association with the disease, and even different alleles within the same gene may be independent. Although the candidate gene approach seems appealing when we have a good hypothesis about the disease, genetic studies often identified genes that were never thought of as candidate genes: a protease inhibitor causing epilepsy, or a channel mutation causing migraine. Therefore we might ask, can we use the power of TDT in detecting alleles that have only a small effect on risk, and combine it with the power of whole genome scanning, which allows identification of new genes not previously thought to be relevant for the disease? In that case, we may need to do TDT tests on all alleles (and there may be dozens in each gene!) on all 100,000 or so genes, an enterprise resulting in over a million data points per individual, with a need to test a few thousand individuals, thus a few billion data points! Statistically, in spite of having to correct for the many different comparisons, this approach was shown to be powerful with a reasonable sample size (Risch and Merikangas 1996). Watson and Akil (current issue) address possible technical solutions that may allow us to imagine such studies in the not too distant future.

In addition, in practice, the number of tests needed may not be quite as large; each population has a history, and it has long been recognized that in populations that started out small not very long ago, for example the Finnish

population (Peltonen et al 1995), many alleles of different loci that are close together tend to have stayed inherited together in a few haplotypes, because they were together on an ancestral chromosome in the relatively recent past. This situation, called linkage disequilibrium (LD), means that there may not be as many possibilities to test; specific alleles of different loci are together on chromosomes in a population more often than expected by chance alone. A specific haplotype, i.e., combination of alleles, rather than single alleles, may then be found associated with disease risk (Templeton et al 1992). In this situation, only a few "representative" alleles of each haplotype may need to be tested for association, reducing significantly the number of tests necessary (Risch and Merikangas 1996). The older and more diverse the original starting population, the shorter the chromosomal regions that are in linkage disequilibrium. Certainly for studies in population isolates, genome scans by association are already possible (Houwen et al 1994). How feasible such association-based genome scans will be in outbred, old populations for complex and heterogeneous diseases is still being debated.

Summary

In summary, at the present time, there is no recipe for how to proceed to identify genes involved in complex diseases such as psychiatric illnesses. Currently, mixed approaches using linkage, association, and linkage disequilibrium are most often performed; for example, once a few chromosomal regions have been tentatively found to be linked, linkage disequilibrium is tested if suitable, and used to narrow down the region, and candidate genes in the most promising regions are scrutinized for possibly functional polymorphic alleles in association studies. Novel ideas, such as the mapping of traits rather than psychiatric illness, have potential, especially if they are less complex in inheritance (Freedman et al current issue). It is also clear that the future will involve more sophisticated statistical analyses, and more automated, large-scale analyses. These will only be fully realized once the human genome has been sequenced, and its diversity has been determined.

Research on Mental Illness in my laboratory is funded by the National Association for Research on Schizophrenia and Affective Disorders (NARSAD), the Nancy Pritzker Network on Depression Research, and the NIAAA.

I thank Adele Barres for help with figures, Huda Akil for many useful discussions, and Michael Boehnke, Scott Stoltenberg, Stanley Watson, James Meador-Woodruff, and Elizabeth Young for critical reading.

This work was presented in December 1997, at the American College of Neuropsychopharmacology Teaching Day Conference on "Complex Genetics: Implications for Psychiatry."

References

Barr CL, Kidd KK (1993): Population frequencies of the A1 allele at the dopamine D2 receptor locus. *Biol Psychiatry* 34:204–209.

Blum K, Noble EP, Sheridan PJ, Montgomery A, Ritchie T, Jagadeeswaran P, et al (1990): Allelic association of human dopamine D2 receptor gene in alcoholism. *JAMA* 263:2055–2060.

Brown DL, Gorin MB, Weeks DE (1994): Efficient strategies for genomic searching using the affected-pedigree-member method of linkage analysis. *Am J Hum Genet* 54:544–552.

Brunner HG, Nelen M, Breakefield XO, Ropers HH, van Oost BA (1993): Abnormal behavior associated with a point mutation in the structural gene for monoamine oxidase A. *Science* 262:578–580.

Chakravarti A (1998): It's raining SNPs, hallelujah? *Nat Genet* 19:216–217.

Collins FS (1995): Positional Cloning moves from perditional to traditional. *Nat Genet* 9:347–350. Collins FS, Guyer MS, Charkravarti A (1997): Variations on a theme: Cataloging human DNA sequence variation. *Science* 278:1580–1581.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, et al (1996): A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380:152–154.

Falk CT, Rubinstein P (1987): Haplotype relative risks: An easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* 51:227–S233.

Freedman R, Adler LE, Leonard S (1999): Alternative phenotypes for the complex genetics of schizophrenia. *Biol Psychiatry* 45:551–558.

Freimer NB, Reus VI, Escamilla MA, McInnes LA, Spesny M, Leon P, et al (1996): Genetic mapping using haplotype, association and linkage methods suggests a locus for severe bipolar disorder (BPI) at 18q22–q23. *Nat Genet* 12:436–441.

Friedman TB, Liang Y, Weber JL, Hinnant JT, Barber TD, Winata S, et al (1995): A gene for congenital, recessive deafness DFNB3 maps to the pericentric region of chromosome 17. *Nat Genet* 9:86–91.

Gelernter J, Goldman D, Risch N (1993): The A1 allele at the D2 dopamine receptor gene and alcoholism. A reappraisal. *JAMA* 269:1673–1677.

Gelehrter TD, Collins FS, Ginsburg D (1998): In: *Principles of Medical Genetics*, 2nd ed. Baltimore: Williams and Wilkins, pp 23–24.

Gershon ES, Badner JA, Goldin LR, Sanders AR, Cravchik A, Detera-Wadleigh SD (1998): Closing in on genes for manic-depressive illness and schizophrenia. *Neuropsychopharmacology* 18:233–242.

Ginns EI, Ott J, Egeland JA, Allen CR, Fann CS, Pauls DL, et al (1996): A genome-wide search for chromosomal loci linked to bipolar affective disorder in the Old Order Amish. *Nat Genet* 12:431–435.

Hebebrand J (1992): A critical appraisal of X-linked bipolar illness. Evidence for the assumed mode of inheritance is lacking. *Br J Psychiatry* 160:7–11.

Houwen RH, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA, et al (1994): Genome screening by searching for shared segments: Mapping a gene for benign recurrent intrahepatic cholestasis. *Nat Genet* 8:380–386.

Kendler KS, Diehl SR (1993): The genetics of schizophrenia: A current, genetic-epidemiologic perspective. *Schizophr Bull* 19:261–285.

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988): Gm3;5,13,14 and type 2 diabetes mellitus: An association in American Indians with genetic admixture. *Am J Hum Genet* 43:520–526.

Lachman HM, Papolos DF, Saito T, Yu YM, Szumlanski CL, Weinshilboum RM (1996): Human catechol-O-methyltransferase pharmacogenetics: Description of a functional polymorphism and its potential application to neuropsychiatric disorders. *Pharmacogenetics* 6:243–250.

Lander E, Kruglyak L (1995): Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nat Genet* 11:241–247.

Lander ES, Schork NJ (1994): Genetic dissection of complex traits. *Science* 265:2037–2048.

Lesch KP, Bengel D, Heils A, Sabol SZ, Greenberg BD, Petri S, et al (1996): Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science* 274:1527–1531.

Malhotra AK, Goldman D (1999): Benefits and pitfalls encountered in psychiatric genetic association studies. *Biol Psychiatry* 45:544–550.

Mitchell P, Mackinnon A, Waters B (1993): The genetics of bipolar disorder. *Aust NZ J Psychiatry* 27:560–580.

Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Manion F, et al (1994): A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). *Science* 265:2049–2054.

National Institute of Mental Health (1998): *Genetics and Mental Disorders*. Rockville, MD: National Institutes of Health.

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengård J, et al (1998): DNA sequence diversity in a 9.7 kb region of the human lipoprotein lipase gene. *Nat Genet* 19:233–240.

Pauls DL, Bailey JN, Carter AS, Allen CR, Egeland JA (1995): Complex segregation analyses of old order Amish families ascertained through bipolar I individuals. *Am J Med Genet* 60:290–297.

Peltonen L, Pekkarinen P, Aaltonen J (1995): Messages from an isolate: Lessons from the Finnish gene pool. *Biol Chem Hoppe Seyler* 376:697–704.

Petit C (1996): Genes responsible for human hereditary deafness: Symphony of a thousand. *Nat Genet* 14:385–391.

Plomin R, Owen MJ, McGuffin P (1994): The genetic basis of complex human behaviors. *Science* 264:1733–1739.

Reich T, Edenberg HJ, Goate A, Williams JT, Rice JP, Van Eerdewegh P, et al (1998): Genome-wide search for genes affecting the risk for alcohol dependence. *Am J Med Genet* 81:207–215.

Risch N (1987): Assessing the role of HLA-linked and unlinked determinants of disease. *Am J Hum Genet* 40:1–14.

Risch N (1989): Linkage detection tests under heterogeneity. *Genet Epidemiol* 6:473–480.

Risch N (1990): Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* 46:222–228.

Risch N, Merikangas K (1996): The future of genetic studies of complex human diseases. *Science* 273:1516–1517.

Risch N, Zhang H (1995): Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* 268:1584–1589.

Roses AD (1998): Alzheimer disease: A model of gene mutations and susceptibility polymorphisms for complex psychiatric diseases. *Am J Med Genet* 81:49–57.

Spielman RS, Ewens WJ (1996): The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 59:983–989.

Templeton AR, Crandall KA, Sing CF (1992): A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* 132:619–633.

Terwilliger JD, Ott J (1992): A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum Hered* 42:337–346.

Van Tol HH, Wu CM, Guan HC, Ohara K, Bunzow JR, Civelli O, et al (1992): Multiple dopamine D4 receptor variants in the human population. *Nature* 358:149–152.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, et al (1998): Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280:1077–1082.

Watson SJ, Akil H (current issue): Gene chips and arrays revealed: A primer on their power and their uses. *Biol Psychiatry*.

Weeks DE, Harby LD (1995): The affected-pedigree-member method: Power to detect linkage. *Hum Hered* 45:13–24.

Weeks DE, Lange K (1988): The affected-pedigree-member method of linkage analysis. *Am J Hum Genet* 42:315–326.

Appendix 1

Glossary

Allele—one of several alternative forms of a gene or an anonymous locus.

APM—affected pedigree member method. A nonparametric linkage method in which alleles that are shared between at least two affected individuals of a family are compared.

ASP—affected sib pair method. A specific form of APM using only affected sibs.

Candidate gene—a gene that might be involved in the disease for biological reasons. Among geneticists, often the meaning is a positional candidate gene, which is a gene that has become a candidate for the disorder because it maps to a chromosomal segment implicated in the disease by linkage.

GRR—genotype relative risk. The increased risk of having a given disorder due to carrying one particular allele.

Haplotype—a combination of closely linked alleles inherited together as a unit.

Heterogeneity—genetic or locus heterogeneity exists if mutations in different genes can cause the same phenotype. This should be contrasted with allelic heterogeneity, which is not discussed in this review and simply means that different mutations in the same gene can cause the same disease; this is common, but does not interfere with genetic mapping.

HRR—haplotype relative risk method. A statistical method for association studies in which the nontransmitted parental alleles are used as controls.

IBD—identity by descent. The identity of two alleles in two individuals because they inherited the same chromosomal segment carrying the allele from a common ancestor.

IBS—identity by state. Identity of two alleles in two people. Identity by state may or may not reflect IBD, which is the more relevant parameter.

LD—linkage disequilibrium. Preferential association of one allele of one locus with a particular allele of another locus more than expected by chance. In the simplest case, a rare disease mutation may be in LD with alleles on nearby loci because the mutation arose only once, on a founder chromosome that carried specific alleles. Those loci close to the mutation have not been separated from the mutation during evolution. Therefore, alleles that were present on the founder chromosome are overrepresented in patients.

Linkage—the close proximity of two loci on a chromosome such that they are separated during meiosis less frequently than expected by chance (50%).

Locus—plural loci: a defined place in the genome, which can be a disease locus, a marker locus, or a gene.

LOD score—the logarithm to base 10 of the likelihood ratio. The likelihood ratio gives the odds favoring linkage at a specific distance θ over the alternative, no linkage. By convention, a LOD score over 3, reflecting an odds ratio of 1000:1, is taken as significant evidence for linkage, a LOD score of -2 (odds 100:1 against linkage) as evidence for excluding linkage.

Microsatellite—see SSLP.

Parametric—model-based. Parametric linkage analysis requires specifying a genetic model, which includes dominance, penetrance, age of onset, and many other parameters. The contrast is nonparametric.

Polymorphism— the occurrence of at least two alleles of a locus in the general population of which the rare allele has a frequency of at least 1%.

RFLP—restriction fragment length polymorphism. A polymorphism that results in an altered pattern of DNA fragments following the action of a specific restriction enzyme.

SNP—single nucleotide polymorphism. Any locus where a simple base change is common in the general population, i.e., polymorphic.

SSLP—simple sequence length polymorphism, also known as microsatellite or STR marker. A locus where a simple sequence (e.g., GT, CAA, or GATA) is repeated in tandem many times, and in which there are multiple alleles of different length, resulting from different numbers of the repeat.

STR—short tandem repeat. See SSLP.

TDT—transmission disequilibrium test. A statistic for association studies. A significant TDT result means that an allele is both associated and linked with a disorder. Parents heterozygous for a predisposing allele transmit that allele more frequently to affected offspring than would be expected by chance, i.e., more than 50%.